Data mining for understanding and improving decision-making affecting ground delay programs

Deepak Kulkarni, Yao Wang and Banavar Sridhar

Section 1. Introduction

The continuous growth in the demand for air transportation results in an imbalance between airspace capacity and traffic demand. The airspace capacity of a region depends on the ability of the system to maintain safe separation between aircraft in the region. The airspace capacity is severely limited by inclement weather. FAA has a national center called Air Traffic Control System Command Center (ATCSCC) that oversees national traffic. Traffic managers at ARTCC collaborate with dispatchers at various Airlines' Operations Center (AOC) to mitigate the demand-capacity imbalance caused by weather. The end result is the implementation of a set of Traffic Flow Management (TFM) initiatives such as ground delay programs, reroute advisories, flow metering, and ground stops.

Data Mining is the automated process of analyzing large sets of data and then extracting patterns in the data. Data mining tools are capable of predicting behaviors and future trends, allowing an organization to benefit from past experience in making knowledge-driven decisions.

In recent years, a number of GDP-related studies using data-mining algorithms have appeared in the literature (Klein, 2009). Since GDP operations are largely developed and carried out without accurate decision support tools in current operations, techniques for modeling the impact of GDP programs prior to operational implementation have been researched in recent years. In (Smith, Sherry, & Donohue, 2008), a decision support capability to predict Aircraft Arrival Rates (AAR) and to determine Ground Delay Program (GDP) program rate and duration based on Terminal Aerodrome Forecast (TAF) weather forecast data using Support Vector Machine (SVM) algorithm, is described. The uses of Ensemble Bagging Decision Tree (BDT), SVM, or Neural Networks (NN) methods to predict the airport capacity and GDP parameters with weather and airport data are introduced in (Wang, 2011) (Wang & Kulkarni, 2011). Despite the past work in this area, there are no published systematic studies seeking to evaluate and predict whether a GDP operation is required or not for days having similar weather and airport conditions.

Data mining algorithms have the potential to develop associations between weather patterns and the corresponding ground delay program responses. If successful, they can be used to improve and standardize TFM decisions resulting in better management of traffic flows on days with reliable weather forecasts. The approach here seeks to develop a set of data mining and machine learning models and apply them to historical archives of weather observations and TFM initiatives to determine the extent to which the theory can predict and explain the observed traffic flow behaviors.

In this study, the major sources of data that were used include: the National Traffic Management Log (NTML) and Aviation System Performance Metrics (ASPM). The data used was from the years 2006 to 2010. The NTML is a unified system developed by the FAA that is used to automate coordination, logging and communication of traffic management initiatives in the NAS. For the purpose of this initial study, the GDP entries in NTML were used as inputs to the data mining algorithms.

A brief overview of the remainder of the paper is as follows. Section 2 discusses ground delay programs. Section 3 provides a high-level overview of data mining techniques that were employed in this study. Section 4 describes the methodology used in the program including metrics and data used in the study. Section 5 presents results. Section 6 is a conclusion.

Section 2. Statistics of Ground Delay Programs

The mission of the FAA's traffic management system to balance traffic demand with system capacity is achieved through a variety of Traffic Mangement Initiatives (TMI) instituted and modified by traffic managers at the regional and national levels. The FAA developed the National Traffic Management Log (NTML) to provide a single system for automated coordination, logging, and communication of TMIs throughout the National Airspace System. Figures below show more detailed GDP event statistics from the data.

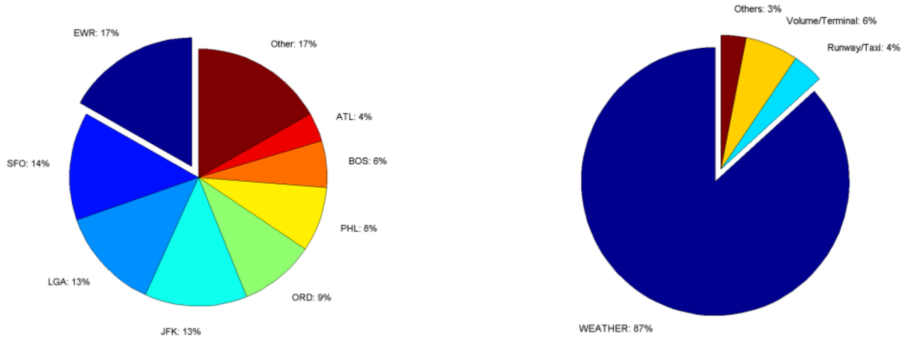

Figure 1: The ratios of GDP counts and GDP causes for the top 8 US airports

Figure 1 displays the ratios in percentage between airport GDP counts and the total NAS GDP counts for the top 8 airports. Fig.1 shows that the most frequent demand-capacity imbalances occurred at the airports in the northeast region of the United

States, such as the three New York-area airports (EWR, LGA, and JFK), Philadelphia (PHL), and Boston Logan International Airport (BOS). The major cause of Ground Delay Programs is weather as demonstrated in Fig.1.

The diverse weather subcategory causes are presented in Figure 2. Details of weather causes for the top 8 airports are provided in Table 1 and 2. Altogether, these data illustrate that the dominated weather causes for GDPs are different at different airports. For example, while close to 90% GDPs at SFO are caused by low ceilings due to marine stratus, wind accounts for about 50% of GDPs at the three New York-area airports, and thunder storms are the major sources of GDP at ATL.

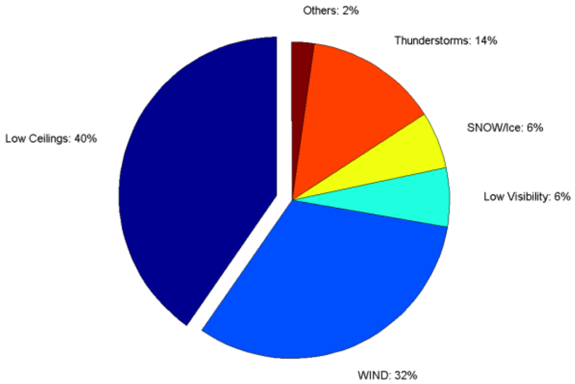

Figure 2: Ratios of the counts between weather subcategories and the total weather GDPs

Airport	*Weather*	*Equipment*	*Center Volume*	*Terminal Volume*	*Runway Taxi*	*Others*
EWR	92%			4%	3%	1%
SFO	96%				3%	1%
LGA	88%	1%		9%	2%	
JFK	78%		1%	17%	2%	2%
ORD	98%	2%				
PHL	91%			1%	8%	
BOS	95%	2%			2%	1%
ATL	96%	4%				

Table 1: Category Percentage Ratio for the Top 8 airports

Airport	*Wind*	*Low Ceilings*	*Low Visibility*	*Rain*	*Fog*	*Snow/Ice*	*Thunder Storms*
EWR	52%	27%	9%	1%		3%	7%
SFO	8%	88%	3%		1%		
LGA	51%	26%	5%	1%		3%	13%
JFK	50%	29%	4%	1%		3%	14%
ORD	29%	25%	8%	6%		14%	15%
PHL	17%	57%	4%	1%	1%	6%	14%
BOS	15%	58%	8%	2%	2%	6%	9%
ATL	5%	37%	9%	1%		3%	45%

Table 2: Weather cause Percentage Ratio for the Top 8 airports

As Newark international airport is an airport that has a very high number of ground delay programs and that contributes significantly to national airspace delays, we initially focused this study at this airport. ASPM is an FAA database containing airport specific data, such as throughput and the weather impacting the airport. Hourly values of wind speed, visibility, ceiling, Instrument Meteorological Conditions (IMC), scheduled arrivals and departures from ASPM data were used to compute input variables in this study.

IMC impacted traffic and wind impacted traffic are two parameters derived from traffic and weather data. As weather impact on the national airspace depends on how many aircraft are impacted by inclement weather, we are using these two metrics to capture the impact of weather on traffic. Wind impacted traffic at an airport was defined as the number of arriving or departing aircraft while wind speed is over 15 knots. Similarly, IMC impacted traffic at an airport was defined as the number of arriving or departing aircraft while there are IMC conditions. Figure 3 below shows histograms of IMC WITI and wind WITI values over a period of five years.

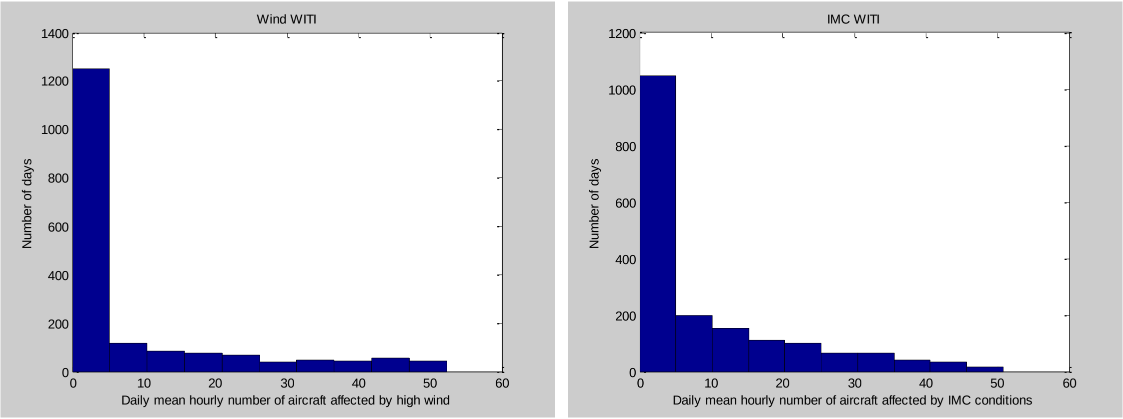

Figure 3. Histograms of IMC WITI and Wind WITI

Daily values of wind WITI and IMC WITI , wind speed, visibility, and ceiling were computed as daily average of hourly values. Values for daily variation in visibility and ceiling are computed as difference between daily maximum and daily minimum values of visibility and ceiling. Daily weather and traffic signature was characterized with following parameters: wind speed, variation in wind speed, visibility, variation in visibility, ceiling, variation in ceiling, Instrument Meteorological Conditions (IMC), scheduled arrivals, IMC impacted traffic and wind impacted traffic.

Principle components analysis of these 10 variables found that the most relevant variables are IMC impacted traffic and wind impacted traffic.

Section 3. Data Mining Overview

In our study, we use three data mining methods: ensemble bagging decision trees (BDT), neural networks (NN), and support vector machine (SVM) learning algorithms.

A. Ensemble Bagging Decision Tree:

Ensemble methods use multiple machine learning models to obtain better predictive performance than what any of its individual constituent members can produce. Bagging is an ensemble method that uses random resampling of a dataset to construct models. In classification scenarios, the random resampling procedure in bagging induces some classification margin over the dataset. Additionally, when bagging is performed in different feature subspaces, resulting classification margins are likely to be diverse, which is essential for an ensemble to be accurate. This method takes into account the diversity of classification margins in feature subspaces to improve the performance of bagging. First, it studies the average error rate of bagging, converts the task into an optimization problem for determining some weights for feature subspaces. Then, it assigns the weights to the subspaces via a randomized technique in classifier construction. Experimental results demonstrate that the ensemble method is robust to classification noise and often generates improved predictions than any single classifier.

B. Neural Networks:

A feed-forward neural network consists of input, hidden and output layers and provides a general framework for representing non-linear functional mapping between a set of input variables and a set of output variables. The output from each layer is connected to the next layer by modifiable weights represented by links between the layers. The weighted outputs from one layer will go through nonlinear sigmoid functions to form the input to the neuron in the next layer. A bias unit is connected to all neurons except the neurons in the input layer. The back-propagation algorithm based on minimizing the output error using a gradient descent method is used for training neural networks. For a NN to have good generalization properties and to avoid over-fitting, the training data should have 5 to 10 times training cases as the weights in NN and it should be statistically representative.

C. Support Vector Machine (SVM)

The Support Vector Machine (SVM), a supervised machine learning algorithm, was invented by Vapnik et al. and has been successively extended by a number of other researchers. Its robust performance with respect to limited, sparse and noisy data is making it widely used in many applications from protein function, and face recognition, to text categorization for classification and regression prediction. The SVM model has also been utilized in airport capacity classification prediction.

When used for binary classification, the SVM algorithm separates a given set of two-class training data by constructing a multidimensional hyper-plane that optimally discriminates between the two clusters. Although SVMs were originally proposed to solve linear classification problems, they can be applied to non-linear decision functions by using the so-called kernel function trick. Adopting this kernel technique, SVM can be utilized to automatically realize a non-linear mapping to a high dimensional space. The hyper plane in the high dimensional space corresponds to a non-linear decision boundary in the input space. A widely used kernel is the Gaussian radial basis function (RBF).

Section 4. Methodology

In some applications, different operators may take different control actions in the presence of similar weather and traffic conditions. Sometimes, the same operator may take different control actions in the presence of similar weather conditions. The reasons for this may be various. Inconsistency may be owing to differing objectives, decision-making styles, or training. The degree of operator decision consistency varies in different regions of the state space. It can be useful to understand the nature of decision inconsistency.

Furthermore, the performance of these data mining methods will vary depending on the state of the system as specified by the observations. The ability of machine learning depends on the consistency of the decision-making process and the availability of the training data in the various regions of the input data state space. Another factor complicating the analysis is lack of clear criterion driving the control actions resulting in different decisions for the same values of the state space. Given the variability in the performance of data mining methods, using a single number to characterize predictive accuracy is not helpful. The paper discusses how to divide data into regions with differing decision consistency and report performance of different data mining methods in the different regions of decision consistency. We will also examine if there is variation in the performance of different data mining methods.

Approach

The general approach adopted in this learning automation work involves the following steps.

1 Division of data into regions of differing decision consistency

2 Comparison of performance of the BDT, NN, and SVM methods in the regions of differing decision consistency

3. Analysis of sensitivity of results to how data is divided into different regions

Metrics used to compare data mining methods

Commonly used metric for evaluating the performance of a data mining method is accuracy which is the proportion of correct predictions. Depending on the situation in which the learnt models are used, it maybe preferable to use a different set of metrics. In predictive analytics, a table of confusion (sometimes also called a confusion matrix), is a table with two rows and two columns that reports the number of false positives, false negatives, true positives, and true negatives. This allows more detailed analysis than mere proportion of correct guesses (accuracy). Accuracy is not a reliable metric for the real performance of a classifier, because it will yield misleading results if the test data set is unbalanced (that is, when the number of samples in different classes vary greatly) and does not reflect actual data for which a model is used. Also, confusion matrix can be particularly important if utility and cost associated with false positives, false negatives, true positives, and true negatives differs significantly.

Observation	**Prediction**	
	Y	N
Y (GDP)	YY	NY
N	YN	NN

In addition to accuracy, we will use critical success index (CSI) and false alarm ratio (FAR) to evaluate performance of different methods. These metrics are defined as follows:

- Critical Success Index (CSI)
  - CSI = YY / (YY + NY + YN).
- False Alarm Ratio (FAR)
  - FAR = YN/(YY + NN)

Section 5. Results

Regions of differing decision consistency

Difficulty of deciding on control action depends on the region of variable space. For example, on clear weather days, most operators would not have any difficulty in concluding that there is no need of weather-caused GDP. Similarly, on really bad weather day, most operators would conclude that there is a need of weather-caused GDP.

Range of values of sum of MC WITI and Wind WITI	Probability of GDP occurrence	Decision consistency	Percent of data
[0, .001)	.14	.86 High	24
[.001, 3)	.23	.77 Medium	9
[3, 11)	.38	.62 Low	17
[11,21)	.61	.61 Low	16
[21,35)	.82	.82 Medium	18
[35, 97)	.92	.92 High	16

Table 3: Segmentation of Data into Multiple Regions

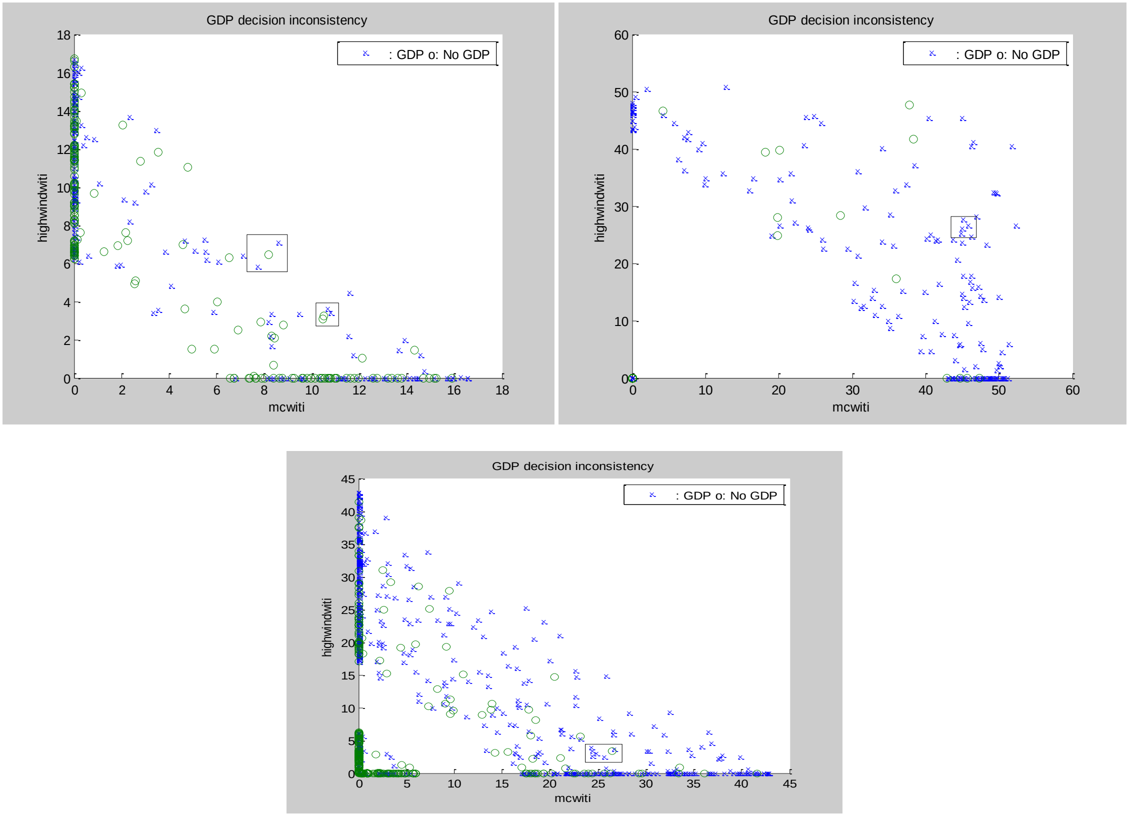

Figure 4: Differing Decision Consistency in Different Regions

Table 1 shows segmentation of data into 6 regions depending on the basis of sum of Wind WITI and MC WITI and probability is computed for corresponding region of data. The number shown in the second column of the table is the percent of cases with GDPs in the particular region of interest. Decision consistency refers to percent of days when the decision was in agreement with the majority decision for the region. As evident in the table above, this value depends on the region of variable space. For example, the first row in the table corresponds to mostly clear weather days. In this case, most operators do not have any difficulty in concluding that there is no need of weather-caused GDP. On the other hand, the last row in the table corresponds to days with the worst weather. In this case, 92% of operators concluded that there is a need of weather-caused GDP. If we examine third and forth rows in the table, we find that about 60% of operator chose to implement GDPs and 40% chose not to. Operators probably need a decision support system in the cases where there seem to be divergence of control actions under the exact same conditions. Given the divergent characteristics of different regions, it would be useful to examine the performance of data mining methods in different regions of data space.

We categorized the six regions as having low, medium or high level of decision consistency and then compared performance of different methods when data has these differing levels of decision consistency.

We found that about 33% of days fall in the category of low decision consistency. About 27% fall in the category of moderate decision consistency and about 40% of days fall in the category of high decision consistency. We also find that performance of data mining methods is better in the region of high decision consistency and is poorer in the region of low decision consistency.

Performance of different methods in the regions of differing decision consistency

Decision Consistency	Percent data	Algorithm	OAR (%)	FAR (%)	CSI (*100)
Low (0.62)	33	NN	66	37	51
		BDT	67	34	52
		SVM	67	31	51
Medium (0.80)	27	NN	79	15	55
		BDT	78	15	52
		SVM	80	13	56
High (0.88)	40	NN	88	19	81
		BDT	87	19	80
		SVM	89	19	82

Table 4: Data Mining Method Performance

Accuracy of these methods varies depending on region of decision consistency. For example, neural network had overall accuracy of 88% in the region of high decision consistency, an accuracy of 79% in the region of medium decision consistency and an accuracy of 66% in the region of low decision consistency. This is not surprising as data mining models can only be as good as the data on which they are trained on.

Utility of data mining methods may vary in different regions of decision consistency. There is probably no need for data mining assistant system in the region of high decision consistency. Data mining methods can be useful in the regions of medium and low decision consistency, but their accuracy is the lowest in the region of low decision consistency.

Sensitivity of results to methods of division into regions of decision consistency

Data segmentation method described in the previous section is not the only method by which we could divide the data. In this subsection, we examine the sensitivity of our conclusions to the method used to divide the data into different parts.

Range of values of sum of MC WITI and Wind WITI	Probability of GDP occurrence	Decision consistency	Percent of data
[0, .001)	.14	.86 High	24
[.001, 6)	.28	.72 Medium	16
[6, 11)	.41	.59 Low	10
[11,17)	.57	.57 Low	10
[17,43)	.80	.80 Medium	30
[43, 97]	.93	.93 High	10

Table 5: Segmentation of Data with Different Thresholds

Analogous to previous section, we divide data into six different regions depending on the sum of MC WITI and Wind WITI. However, we used different thresholds in our case. Next, we characterized the six regions as having low, medium or high level of decision consistency. We found that about 20% of days fall in the category of low decision consistency. About 46% fall in the category of moderate decision consistency and about 34% of days fall in the category of high decision consistency. The percent of data that is in region of low decision consistency with this segmentation differs from that in the previous section. So, percent numbers are sensitive to how segmentation is done and how high, medium and low levels of decision consistency are defined. Table below shows the performance of data mining methods in the different regions. Different data mining methods have similar performance in different regions. We also find that performance of data mining methods is better in the region of high decision consistency and is poorer in the region of low decision consistency. For example, Neural network accuracy (OAR) is 87% in the region of high data consistency and it drops to 65% in the region of low data consistency.

Decision Consistency	Percent data	Algorithm	OAR (%)	FAR (%)	CSI (*100)
Low (0.58)	20	NN	65	39	50
		BDT	68	32	53
		SVM	66	35	51
Medium (0.77)	46	NN	79	16	55
		BDT	77	17	51
		SVM	79	20	58
High (0.88)	34	NN	87	28	82
		BDT	87	25	82
		SVM	88	27	84

Table 6: Performance of Data Mining Methods on Second Set of Regions

Sensitivity of the method of variable set used

Table below shows the performance of different data mining methods when input parameters include AAR as well. The purpose of using AAR is two-fold. First of all, depending on the purpose of analysis, it is possible that AAR is an input that could be used. For example, if the purpose of the analysis post-operations analysis, AAR information is readily available and one may want to use it as a part of analysis. Secondly, we may want to check whether general conclusions of this study of valid in the presence of different set of variables.

Decision Consistency	Algorithm	OAR (%)	FAR (%)	CSI (*100)
Low (.58)	NN	77	27	64
	BDT	76	25	63
	SVM	77	24	63
Medium (.77)	NN	84	12	64
	BDT	82	12	61
	SVM	83	18	65
High (.88)	NN	87	24	82
	BDT	88	23	83
	SVM	88	25	84

Table 7: Performance of Data Mining Methods With AAR Included As Input

Again, the general conclusions of still valid. Different data mining methods have similar performance in different regions. We also find that performance of data mining methods is better in the region of high decision consistency and is poorer in the region of low decision consistency.

Section 6. Conclusion

Difficulty of deciding on control action depends on the region of variable space. Weather signature on different days can categorize days into days with little decision difficulty, days with moderate decision difficulty and days with high decision difficulty. This paper reported performance of different data mining methods in the three regions of decision difficulty. Not surprisingly, data mining methods have the best performance in the region of little decision difficulty and have the poorest performance in the region of most decision difficulty. In applications where data mining methods have differing performance in differing regions, it would be more useful to characterize the region specific performance instead of characterizing performance by a single parameter.

Also, there is probably not need for data mining assistant system in the region of small consistency. Therefore, operators may find decision support systems to be most useful in the regions of moderate or low decision difficulty. Also, organizations may want to examine decision making processes that are used in these regions to see how much subjectivity exists. Thus, it may be useful to segment data and identify the regions of low and moderate decision consistency.

Finally, we also found that there was not significant variation in the performance of different data mining methods for this particular problem. The fact that different mining methods show no significant variation also provide further confidence in the results of data mining methods.

Bibliography

Breiman, L. (1996). Bagging Predictors . *Machine Learning , 24* (2), 123-140.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology , 2* (21), 1--27.

Chang, C.-C. C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology , 2* (21), 1--27.

Foresee, F. a. (1997). Gauss-Newton approximation to Bayesian regularization. *International Joint Conference on Neural Networks*, (pp. 1930-1935).

Foresee, F., & Hagan, M. T. (1997). Gauss-Newton approximation to Bayesian regularization. *International Joint Conference on Neural Networks*, (pp. 1930-1935).

Jain, A., & Dubes, R. (1988). *Algorithms for Clustering Data.* Prentice Hall.

Klein, A. S. (2009). Weather Forecast Accuracy: Study of Impact on Airport Capacity and Estimation of Avoidable Costs,. *Eighth USA/Europe Air Traffic Management Research and Development Seminar .*

Scholkopf, B. A. (2002). *Learning with kernels, support vector machines, regularization, optimization and beyond.* Cambridge: MIT Press.

Scholkopf, B., & Smola, A. J. (2002). *Learning with kernels, support vector machines, regularization, optimization and beyond.* Cambridge: MIT Press.

Smith, D. A. (2008). Decision Support Tool for Predicting Aircraft Arrival Rates, Ground Delay Programs and Airport Delays from Weather Forecasts. *ICRAT.* Fairfax, VA.

Smith, D. A. (2008). Decision Support Tool for Predicting Aircraft Arrival Rates, Ground Delay Programs, and Airport Delays from Weather Forecasts, . *Proceedings International Conference on Research in Air Transportation.* Fairfax, VA. .

Smith, D. A., Sherry, L., & Donohue, G. (2008). Decision Support Tool for Predicting Aircraft Arrival Rates, Ground Delay Programs and Airport Delays from Weather Forecasts. *ICRAT.* Fairfax, VA.

Vapnik, V. (1998). *Statistical Learning Thoery.* New York: John Wiley and Sons.

Wang, Y. (2011). Prediction of weather impacted airport capacity using ensemble learning,. *Digital Avionics System Conference.*

Wang, Y., & Kulkarni, D. (2011). Modeling Weather Impact on Ground Delay Programs. *SAE International Journal of Aerospace , 4* (2), 1207-1215.